Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 47-55 recibido 02-05-2011 aceptado 27-05-2011

Resolución de Correferencia de Nombres de Persona para Extracción de Información Biográfica *

Personal Name Coreference Resolution for Biographical Information Extraction

Marcos Garcia
Centro de Investigación en Tecnoloxías da Información (CITIUS)
Univ. de Santiago de Compostela
marcos.garcia.gonzalez@usc.es

Pablo Gamallo
Centro de Investigación en Tecnoloxías da Información (CITIUS)
Univ. de Santiago de Compostela
pablo.gamallo@usc.es

Resumen: Los sistemas de extracción de información necesitan un procesamiento previo que reconozca, entre otras cosas, elementos correferenciales tales como las variantes de nombres propios. El presente artículo tiene dos objetivos: por un lado, describe los principales tipos de correferencia de nombres de persona encontrados en textos enciclopédicos y periodísticos en castellano. Por otro lado, presenta un algoritmo que resuelve satisfactoriamente la mayor parte de los casos descritos. El sistema, que no necesita *corpus* de entrenamiento, permite unificar las variantes de nombres de persona que aparecen en un texto, mejorando así tareas como la extracción de información biográfica.
Palabras clave: correferencia de nombres de persona, extracción de información.

Abstract: Information extraction systems need a previous processing step in order to recognize coreferential elements, such as personal name variants. This paper has two aims: the first is to describe the main types of personal name coreference found in encyclopedic and journalistic texts in Spanish. Furthermore, we introduce an algorithm that solves most coreferential links between personal name variants succesfully. The system, which does not need a training *corpus*, unifies the coreferential elements found in a text, thereby improving tasks like biographical information extraction.
Keywords: personal name coreference, information extraction.

1. Introducción

En los últimos años, la cantidad de información que podemos encontrar en Internet está aumentando exponencialmente. A consecuencia de esto, el interés en obtener datos semánticos automáticamente también se ha visto incrementado.

La extracción de información biográfica tiene como objetivo crear de manera automática grandes repositorios que contengan información semántica estructurada acerca de personajes públicos: ocupación, fecha y lugar de nacimiento y/o muerte, obra, etc. Por ejemplo, de una frase como "Linus Benedict Torvalds (Helsinki, Finlandia, 28 de diciembre de 1969), es un ingeniero de software finlandés.", el objetivo sería extraer la siguiente información:

- Linus Benedict Torvalds *Lugar de Nacimiento* Helsinki, Finlandia
- Linus Benedict Torvalds *Fecha de Nacimiento* 28/12/1969
- Linus Benedict Torvalds *Ocupación* ingeniero de software
- Linus Benedict Torvalds *Nacionalidad* finlandés

Esta información puede ser utilizada posteriormente en sistemas de búsqueda de respuestas (Mann, 2002), en procesos de recuperación de información (Wan et al., 2005) o en ampliación de ontologías (Suchanek, Ifrim, y Weikum, 2006), por ejemplo.

Una buena parte de las estrategias de extracción de información de fuentes no estructuradas consiste en (i) obtener oraciones que contengan pares de entidades potencialmente representativos de una relación semántica

* Este trabajo ha sido subvencionado por el Ministerio de Ciencia e Innovación, con cargo al proyecto con referencia FFI2010-14986.

(Nombre_de_Persona - *Fecha_de_Nacimiento* - Fecha) y (ii) decidir si realmente el par contiene esa relación en la oración extraída.

Un problema habitual que surge en este proceso está relacionado con la variabilidad que suelen presentar los nombres de persona (NPer). Así, en un único texto no es extraño que una misma persona aparezca referida de varias maneras: "Miguel de Cervantes", "Miguel de Cervantes Saavedra", "M. de Cervantes", "Miguel", "El manco de Lepanto", "Cervantes" (que a su vez puede referirse a una localidad), etc., al lado de otras entidades como "Calle Miguel de Cervantes" ó "Universidad Miguel de Cervantes".

La resolución de la correferencia existente entre las diferentes variantes de NPer que aparecen en un texto es fundamental para la obtención y estructuración de información. Aparentemente trivial, la desambiguación de algunos de estos casos presenta varias dificultades. Aun así, la bibliografía relativa a este problema no es abundante.

El objetivo de este artículo consiste, por un lado, en describir los principales tipos de correferencia de NPer que existen en textos escritos en castellano. Por otro lado, presentamos un algoritmo de resolución de este tipo de correferencia que no necesita *corpus* de entrenamiento. El sistema, evaluado sobre dos *corpora* (periodístico y enciclopédico), permite unificar los diferentes nombres de persona correferenciales, permitiendo así mejorar subsiguientes tareas de fusión y desambiguación de información obtenida de varios documentos. Los experimentos son realizados sobre *corpora* en castellano, aunque el sistema es fácilmente adaptable a otras lenguas.

Además de esta sección introductoria, la Sección 2 hace un breve recorrido sobre trabajos relacionados. En la Sección 3, presentamos los principales tipos de correferencia de NPer. La Sección 4 describe el algoritmo de resolución. Finalmente, la Sección 5 muestra los experimentos realizados y la Sección 6 concluye este trabajo.

2. Trabajo relacionado

Uno de los procesos previos a la resolución de correferencia entre nombres propios consiste en la propia detección y etiquetación de estos. Así, los primeros trabajos que se dedicaron a este tipo de correferencia se centraron en resolver ambigüedades estructurales (como la detección de de las fronteras de los nombres propios: "Universidad Complutense de Madrid" *vs* "Museo Reina Sofía *de Madrid*") y semánticas (entre diferentes tipos de entidades: personas, localidades, organizaciones u otras) (Mani et al., 1993; Wakao, Gaizauskas, y Wilks, 1996; Wacholder, Ravin, y Choi, 1997).

De modo más detallado, Kim y Evens (1996) diseñaron un algoritmo que, después de recorrer y guardar todos los nombres propios de un texto (excepto las *keywords* y palabras funcionales: Sr., Mr., Coorp., etc.), realiza diferentes tipos de *matching* entre ellos. De este modo consiguen discriminar nombres que pueden pertenecer tanto a organizaciones como a personas: "Cray Computer" (organización) *vs* "Seymour Cray" (persona).

La popularización de los sistemas de reconocimiento y/o clasificación de entidades con nombre (NER/NEC) resolvió gran parte de las ambigüedades señaladas, por lo que el interés por la resolución de este tipo de correferencia disminuyó. Así, es habitual asumir que, en un mismo texto, no existen ambigüedades entre NPer, ya que el autor tiende a evitar ese tipo de confusiones durante la redacción (Fleischman y Hovy, 2002).

De todas formas, dependiendo del tipo y longitud del texto, pueden ser muchas las variantes que tome un nombre propio en un mismo documento. Bontcheva et al. (2002) presentan, dentro de un sistema más extenso de resolución de correferencia, un conjunto de reglas para establecer relaciones entre elementos con el mismo referente: *matching* de diferentes tokens en función de su posición, expansión de siglas y acrónimos, listas de equivalencias, etc.

Las aproximaciones a la resolución total de correferencia (no sólo entre nombres propios o de persona) han sido objeto de varias evaluaciones (*MUC-7* o *SemEval 2010*). Dada su complejidad, este tipo de tareas requieren un mayor número de recursos, tales como *corpora* anotados, analizadores sintácticos, etc. (Mitkov et al., 2000; Recasens y Hovy, 2010).

Otro tipo de trabajos que utilizan algunas de estas técnicas son los que realizan *personal name matching*, proceso que consiste en encontrar variantes de nombres propios provocadas por diferencias ortográficas, fonéticas, entre idiomas, etc. (Gadafi, El Gaddafi, Qaddafi, Al-Gathafi, Kaddafi, etc.) (Cohen, Ravikumar, y Fienberg, 2003).

Por último, cabe destacar que la resolución de correferencia entre los nombres propios de un mismo documento es también útil para la posterior desambiguación de entidades entre múltiples documentos, donde "Manuel Rivas" podría referirse a un escritor gallego, a un ajedrecista de Jaén, o a un político chileno, por ejemplo (Mann y Yarowsky, 2003).

3. Correferencia de nombres de persona

En esta sección describimos los principales tipos de correferencia entre nombres de persona encontrados en un mismo texto en castellano, así como las dificultades que su desambiguación puede conllevar. Los diferentes ejemplos se corresponden con casos reales extraídos de la Wikipedia en español.

Téngase en cuenta que en la detección y resolución de correferencia entre NPer entran en juego varios factores: además del propio algoritmo de resolución, este puede requerir recursos externos de conocimiento (listas de variantes, *keywords*, equivalencias, etc.) o sistemas de reconocimiento y clasificación de nombres propios.

3.1. Estrategias de resolución

En primer lugar, hemos definido dos tipos fundamentales de aproximaciones a la resolución de correferencia entre NPer, en función de los objetivos o de los recursos utilizados:

Correferencia con Foco: este tipo de estrategias son aplicadas sobre textos cuyo foco se centra en una única persona, como son los artículos enciclopédicos o biográficos. Así, esta aproximación se centra en encontrar las variantes que el nombre principal tiene en el texto, ignorando la correferencia que pueda existir entre otras entidades. Estas técnicas, básicamente de *matching*, son rápidas e eficaces en este tipo de textos, aunque no son fácilmente aplicables en documentos cuyo foco no sea claro y, si existen, no establecen referencias entre nombres diferentes de este.

Correferencia Completa: a diferencia de la resolución de correferencia *con foco*, este tipo de estrategias intentan establecer las relaciones entre todos los NPer que compartan algún referente en un texto. En estos casos, la existencia de un mayor número de candidatos y de referentes implica también un aumento de la ambigüedad.

Si bien es cierto que los textos enciclopédicos contienen grandes cantidades de información biográfica sobre una persona en particular, en estos documentos podemos encontrar datos relevantes sobre otras personas. Además, documentos de otras tipologías suelen tener información precisa acerca de diferentes personalidades. Por esta razón, tanto la descripción de los tipos de correferencia como el algoritmo presentado tienen como base la resolución de correferencia *completa*.

3.2. Tipos de correferencia

Antes de presentar los diferentes tipos de correferencia entre NPer que hemos definido, es preciso referir que en el presente trabajo nos centramos en las relaciones entre entidades que comparten alguna relación léxica ("Jacques Chirac" y "Presidente Chirac"), y no en aquellas correferencias entre entidades más generales ("President de la Generalitat de Catalunya" - "Artur Mas").

Con todo, los casos tratados en el presente artículo tienen en cuenta la *transitividad*, es decir, aquellas relaciones entre dos elementos correferenciales que son establecidas a través de un tercero: "Camilo José" y "Cela" no contienen una relación léxica, pero pueden ser definidos como correferenciales si en el mismo texto se encontrase un nombre propio como "Camilo José Cela y Trulock".

3.2.1. Inclusión

Este tipo de correferencia se da entre una variante de un nombre propio cuya forma se incluye en una mayor con la cual comparte referente ("Lennon" y "John Lennon"). Para resolver estos casos, debe tenerse en cuenta que, en nombres con más de dos tokens ("John Winston Ono Lennon"), la forma abreviada puede contener uno o más elementos, adyacentes o no en la variante mayor: "John", "Lennon", "John Lennon", etc.

Además, la utilización de uno u otro apellido puede variar en función de la lengua o cultura del texto: mientras en castellano suele utilizarse el primer apellido como principal ("Gaspar Llamazares Trigo": "Llamazares", y no "Trigo"), en otras culturas lo más habitual es referirse a una persona a través de su último apellido ("Fernando António Nogueira de Seabra Pessoa": "Pessoa" o "Fernando Pessoa"). Con todo, existen casos en los que el uso de un apellido concreto ayuda a desambiguar a la persona, y que por lo tanto difieren de las *reglas* generales de la

lengua/cultura: "José Luis Rodríguez Zapatero": "Zapatero".

La mayor parte de los casos de inclusión pueden resolverse con técnicas simples de *matching* entre los tokens de las dos variantes en causa. Aun así, existen casos como "John Lennon" y "John Jack Lennon" (abuelo del primero) o "George Herbert Walker Bush" y "George Walker Bush" (padre e hijo), que serían mal analizados con esta estrategia.

Determinadas clases de palabras deben ser tenidas en cuenta, dada su importancia en la configuración de los NPer:

1. Palabras vacías, con las cuales el *matching* resulta inefectivo: la búsqueda de variantes de "Javier de la Torre" debe realizarse solamente a través del primer y último elemento.
2. Fórmulas de tratamiento (o *keywords* positivas), como "Sr.", "Doña", etc.
3. *keywords* negativas, que bloquean la relación entre dos NPer: "John F. Kennedy" - "John F. Kennedy **Jr.**".

Por último, y en función de la utilización o no de un sistema NEC (y de su precisión), la resolución de la correferencia por inclusión debe tener en cuenta casos como "Universidad Fernando Pessoa" - "Estatua de Fernando Pessoa" - "Antología de Fernando Pessoa", etc. que pueden ser erróneamente clasificados como variantes de "Fernando Pessoa".

3.2.2. Abreviatura

Un tipo diferente de correferencia entre NPer se da entre entidades en las cuales uno o más de sus constituyentes han sido reducidos a su abreviatura. Este tipo de casos también presenta un gran variabilidad, ya que (i) puede haber una o más formas abreviadas, (ii) pueden ser formas iniciales, intermedias, o finales y (iii) las abreviaturas pueden ir seguidas de punto o no.

Ejemplos de este tipo correferencia, con abreviaturas inicial, intermedia y total son los siguientes: "John Fitzgerald Kennedy": "John F. Kennedy", "J. F. Kennedy", "J.F.K.", "JFK" (que también puede ser el nombre de un aeropuerto y de una película).

Un ejemplo de nombre propio cuyo elemento abreviado es el último sería: "Camilo J.": "Camilo José Cela". Recuérdese que algunos de estos casos requieren transitividad para su resolución ("Camilo J." - "Cela").

Debemos tener en cuenta que, del mismo modo que el anterior tipo de correferencia, pueden existir excepciones que no serían bien analizadas con simples reglas de *matching* y de expansión de abreviaturas: "George H. W. Bush" (padre) y "George W. Bush" (hijo).

3.2.3. Hipocorísticos y Diminutivos

La correferencia de NPer a través de hipocorísticos y diminutivos es un caso frecuente en textos de diversas lenguas. A pesar de que el uso más habitual de este tipo de palabras no se encuentra en textos enciclopédicos o periodísticos, existen casos de hipocorísticos y diminutivos que, dada su frecuencia de uso, se pueden considerar lexicalizados.

Así, nombres como "José Blanco" se encuentran referenciados como "Pepe Blanco" o como "Pepiño". Este último ejemplo muestra que los diminutivos pueden variar en función de la comunidad lingüístico-cultural a la que pertenece el nombre referenciado, o en la que es referido. Así, "Eva Duarte" es nombrada como "Evita" y "Anthony Charles Lynton Blair" como "Tony Blair".

Hay que tener en cuenta que la relación entre los hipocorísticos y los NPer no siempre son unívocas. El hipocorístico "Berto" puede tener como referente "Alberto", "Roberto", "Norberto", "Heriberto", etc., siendo esta característica más acentuada en lenguas como el inglés, donde la variante "Bert" puede corresponderse con más de 10 nombres masculinos y femeninos. Por lo tanto, la simple recopilación de nombres propios e hipocorísticos puede no resolver la ambigüedad entre elementos correferenciales.

3.2.4. Apodos

Un caso similar al anterior es aquél en el que uno de los elementos correferenciales es un apodo del nombre propio.

Entre los apodos pueden encontrarse también diminutivos ("Ronaldo de Assis Moreira": "Ronaldinho"), pero también otras características que no encontramos en el nombre propio: "Ronaldinho *Gaúcho*" o "Elvis Aaron Presley": "El Rey del Rock".

La relación entre los apodos y su nombre propio correferencial suele ser más unívoca que en el caso de los hipocorísticos, aunque existen excepciones (que normalmente no son ambiguas en un único texto): "Juninho", que puede referirse, entre otros a los futbolistas "Juninho Pernambucano" y "Juninho Paulista".

Esto implica que la utilización de listas de variantes entre nombre propio y apodo es más efectiva que en el caso de los hipocorísticos. Con todo, cabe destacar que, salvo casos lexicalizados ("Edson Arantes do Nascimento": "Pelé"), el uso de los apodos en textos enciclopédicos y/o periodísticos sólo suele utilizarse en la presentación (p. e. "*Nombre Propio*, también conocido como *apodo*"), por lo que no aparece en muchos más contextos que potencialmente contengan información relevante para ser extraída.

3.2.5. Foco/Conocimiento General

Como hemos dicho en la Sección 2, es habitual asumir que, en un mismo documento, no encontraremos referencias parciales a individuos que compartan apellido (o nombre). Así, un redactor no incluirá la forma "Lennon" en un documento en el que coocurran nombres como "John Lennon" y "Alfred Lennon".

Sin embargo, este tipo de casos son muy comunes en dos contextos: (i) en documentos enciclopédicos, cuyo foco o nombre principal está marcado en el propio título o encabezamiento, y (ii) en textos sobre personajes que, socialmente, son más relevantes que otros con los que comparten nombre o apellido.

Ejemplos del primer caso pueden ser los ya citados "John Lennon" y "Alfred Lennon", que aparecen en un mismo documento al lado de "Lennon". Lo mismo ocurre en artículos como el de "George Bush", donde es habitualmente referido con "Bush", y en el que también aparece su esposa "Laura Bush".

Los dos anteriores ejemplos pueden ser considerados también como personajes relevantes: "Lennon" y "Bush" son habitualmente utilizados para referirse a "John Lennon" y a "George Bush", debido, p. e., a su mayor relevancia histórica y/o social.

Con todo, existen ejemplos de este tipo (el segundo de los casos citados) que aparecen en documentos cuyo foco no recae sobre ellos, y que su conocimiento viene dado únicamente por un conocimiento externo. Así, en un único texto sobre "Miguel de Cervantes", encontramos varias referencias a "Lope", al lado de los nombres "Lope de Vega", "Lope de Rueda" y "Lope de Figueroa" (algunas tan próximas entre sí que incluso un lector podría tener dificultades en su desambiguación). La resolución de este tipo de casos requiere recursos externos de conocimiento general (y dinámico, ya que este varía dependiendo del contexto y del tiempo).

Finalmente, existen otros casos problemáticos de correferencia, que se dan en metatextos. Algunos documentos (principalmente referentes a escritores), incluyen extractos de obras, referencias a personajes, etc., cuyo nombre o apellido pueden coincidir con otros NPer (externos a la obra referida), dificultando así la resolución. Cabe señalar, con todo, que no son casos muy frecuentes, y que su incidencia en tareas de extracción de información es probablemente reducida.

4. Descripción de los algoritmos

En esta sección describiremos los dos algoritmos que, junto a un *baseline*, han sido utilizados para la resolución de correferencia de nombres de persona. Los sistemas son basados en reglas, por lo que no necesitan un *corpus* de entrenamiento previamente etiquetado, e independientes de una lengua particular. Sólo los recursos externos son dependientes, pero se pueden extraer fácilmente de manera automática.

***Baseline*:** como primer modelo de comparación, hemos definido un *baseline* que extrae únicamente los casos en los que ocurre el nombre propio en su forma enciclopédica, por lo que no se crean *links* de correferencia (similar a *all singletons* de Recasens y Hovy (2010)).

Así, los nombres incluidos en un artículo se comparan con una lista, previamente extraída, de títulos de artículos (que pertenezcan a categorías relativas a personas) de la Wikipedia. A modo de ejemplo, de las formas "John Lennon", "Lennon", "John Winston Ono Lennon" o "Discografía de John Lennon", únicamente la primera sería extraída, por ser la forma más habitual que coincide con el título del artículo. Nótese que esta forma enciclopédica es la utilizada normalmente en los "pares semilla" aplicados para buscar patrones de extracción ("John Lennon" - "músico", etc.).

Las únicas herramientas necesarias para aplicar este modelo son un tokenizador y un NER.

***Partial-match*:** esta estrategia de correferencia de NPer está basada en *matching* parcial. Para ello, además del tokenizador y el NER, aplicamos un clasificador de entidades (NEC), que las etiqueta como persona, organización, localidad u otros (*misc*).

El algoritmo hace lo siguiente con cada uno de los nombres etiquetados como "persona" por el NEC:

1. Establece relaciones entre los NPer con la misma forma.
2. Cada NPer es dividido en tokens, ignorando las palabras vacías que puedan contener (<de>, <la>, etc.).
3. Los tokens de un NPer (A) son comparados con los tokens de otro (B), estableciéndose una relación de correferencia si algún token de A existe en B, excepto:
   - En casos en los que un nombre simple ("John") pueda ser correferencial de más de un nombre compuesto ("John Lennon" o "John F. Kennedy"), donde la relación se establece entre el nombre compuesto anterior más próximo.
   - En casos en los que el número de coincidencias entre los tokens de A y B es menor que entre B y C: (A) "Fernando González Ochoa"; (B) "José González" y (C) "José González Torres".

NP-Var: para aplicar el siguiente algoritmo es necesario obtener (además de un conjunto de palabras vacías) las siguientes listas de palabras:

- *Trigger-words* de Persona (*tw-p*): palabras que indican que la entidad es una persona (presidente, cónsul, etc.). Son obtenidas automáticamente a través de métodos utilizados en el desarrollo de sistemas NEC, o extraídos directamente de recursos libres (Carreras, Màrquez, y Padró, 2003).
- *Trigger-words* de Localidades, Organizaciones (y otras) (*tw-block*): calle, museo, instituto, gobierno, obra, etc. Obtenidas también de sistemas NEC libres y aumentadas automáticamente de la siguiente manera: se extraen todos los nombres propios de un *corpus* de gran tamaño (en nuestro caso, la Wikipedia), y se seleccionan aquellos cuyo primer token sea una palabra de diccionario (p. e. museo) y su segundo token sea un NPer (estos últimos previamente extraídos de modo automático de la Wikipedia). La lista es posteriormente filtrada con nombres de pila de persona, para evitar entradas que, siendo formas de diccionario, sean también nombres de pila (p. e. "Rosa").

Una vez obtenidos estos recursos, para aplicar el algoritmo, el texto fuente es también analizado con un tokenizador, un NER y un NEC.

El primer paso consiste en decidir si el documento analizado tiene como foco un NPer. Si el título o encabezamiento del documento se corresponde con un NPer (p. e., artículo enciclopédico), este es seleccionado como foco. Si no, (i) se selecciona el NPer de un sólo token más frecuente del documento; (ii) se verifica si este nombre forma parte de uno de los n NPer compuestos más frecuentes del texto (donde n ha sido definido empíricamente como 2); (iii) si es así, se define el nombre compuesto más frecuente como foco, salvo que exista otro NPer compuesto de frecuencia similar ($>= 0,6$) que también contenga el nombre simple. El grado de similaridad entre el primer y el segundo candidato ha sido establecido en $0,6$ después de haber probado diferentes valores.

Si el documento tiene como foco un NPer, todos las apariciones simples de cada uno de los tokens del nombre-foco se etiquetan como correferencial de este.

Tanto si se ha encontrado foco como si no, se realiza *pattern-matching* entre los NPer y NP*misc*, estableciendo correferencias entre ellos. Nótese que el algoritmo utiliza también las entidades *misc* (que no son persona, localidad u organización), debido a que pueden corresponderse con NPer mal etiquetados.

A continuación, cada NPer y NP*misc* encontrado desde el inicio (A), es comparado con los anteriores (B):

1. Si A sólo tiene un token, se considera correferencial de B si B contiene ese token (y A no es una *tw-block*).
2. Si A tiene más de un token, se considera correferencial de B si B contiene a A (ignorando las palabras vacías y las *tw-p*), salvo que A o B terminen en una *keyword* negativa ("Jr."), no presente en el otro nombre.
3. Si A tiene un token de un sólo carácter (o una letra y un punto), que no sea el último, se selecciona el primer carácter

de todos los tokens de A y B (salvo el último, las palabras vacías y *tw-p*), y se comparan: si B contiene a A, se consideran correferenciales.

4. Si el token de un único carácter es el último, se realiza la misma operación, pero incluyendo todos los tokens (excepto palabras vacías y *tw-p*).

5. Después de recorrer todos los nombres del documento, el algoritmo vuelve al principio. Todos aquellos NPer y NP*misc* para los que no ha sido encontrado un correferente son comparados con el NPer y NP*misc* siguiente más próximo, realizando las mismas operaciones (nótese que ahora A y B se invierten).

Finalizada la ejecución, son obtenidas todas las variantes de cada uno de los nombres propios, y su lema es substituido por una forma canónica (el nombre más largo y una identificación numérica).

5. Experimentos

Para conocer el funcionamiento de los sistemas presentados, utilizamos un *corpus* de evaluación compuesto de textos enciclopédicos y periodísticos. Para crearlo, escogimos aleatoriamente 12 noticias del periódico español El País (secciones "España" e "Internacional") y 10 artículos de la versión española de la Wikipedia que perteneciesen a la categoría "Escritores en Español". Ninguno de los textos había sido utilizado durante el desarrollo de los algoritmos. Las relaciones de correferencia fueron anotadas y revisadas manualmente, encontrándose un total de 390 *links* entre NPer (141 en las noticias periodísticas y 249 en los artículos enciclopédicos).

El preprocesamiento del *corpus* (tokenización, NER y NEC) fue realizado con FreeLing (Padró et al., 2010).

Para evaluar el funcionamiento de los modelos, utilizamos dos medidas diferentes:

1. **Positiva** (basada en la evaluación MUC (Vilain et al., 1995)): esta evaluación se realiza sobre los casos positivos. Así, la precisión es calculada dividiendo el número total de *links* (entre entidades correferenciales) correctos entre la suma de los *links* correctos y los incorrectos. El *recall* se obtiene dividiendo el número de *links* establecidos correctamente entre el número total de *links* existente en el *gold-standard*.

2. **BLANC** (Recasens y Hovy, 2011): este tipo de evaluación tiene en cuenta no sólo los *links* de correferencia existentes (*coreference*), sino también los pares de entidades que no tienen una relación de correferencia (*non-coreference*). Así, la precisión y el *recall* son calculados de manera independiente sobre los casos de correferencia y de no-correferencia. Finalmente, cada una de las medidas BLANC (Precision, *recall* y f-score) se corresponde con una media de los valores calculados independientemente.

Positiva		**Wikip.**	**El País**	**Total**
Basel.	P	100	100	100
	R	27,7	29,1	28,2
	F	43,4	45,1	**44**
P-Match	P	79,4	95,6	86
	R	89,9	93,6	91,2
	F	84,3	94,6	**88,5**
NP-Var	P	94,5	98,4	96
	R	89,9	91,4	90,5
	F	92,2	94,8	**93,1**

Tabla 1: Precisión, *Recall* y *f-score* (medida *positiva*) de los tres sistemas evaluados en los *corpora* extraídos de la Wikipedia y El País.

BLANC		**Wikip.**	**El País**	**Total**
Basel.	P	99,7	99,4	99,6
	R	63,9	64,5	64,2
	F	71,5	72,3	**71,9**
P-Match	P	90,2	98,5	94,4
	R	94,2	99,2	96,7
	F	92,1	98,9	**95,5**
NP-Var	P	97,8	99,5	98,7
	R	94,4	95,4	94,9
	F	96,1	97,4	**96,7**

Tabla 2: Precisión, *Recall* y *f-score* (medida *BLANC*) de los tres sistemas evaluados en los *corpora* extraídos de la Wikipedia y El País.

Las Tablas 1 y 2 contienen los resultados de la evaluación de los dos sistemas presentados (junto al *baseline*) utilizando las medidas *Positiva* y BLANC, respectivamente.

La precisión del sistema *baseline* es del 100 %/99,6 %, dado que sólo reconoce los

Positiva	**Wikipedia**	**El País**	**Total**
Prec.	95,7	98,5	96,7
Rec.	97,4	98,5	97,8
F-score	96,5	98,5	97,2

BLANC

Prec.	98,6	99,6	99,1
Rec.	99,4	99,6	99,5
F-score	99	99,6	99,3

Tabla 3: Precisión, *recall* y *f-score* del sistema NP-Var asumiendo un preprocesamiento óptimo.

NPer que coinciden totalmente con el nombre enciclopédico; esto conlleva valores bajos de *recall* y, por lo tanto, también de *f-score*.

Entre los dos sistemas comparados (*partial-match* y NP-Var), NP-Var obtiene mejores resultados, si bien es cierto que los valores de *recall* de *partial-match* son superiores en textos periodísticos. En este sentido, cabe destacar que los casos de correferencia son resueltos con más precisión en el *corpus* de El País que en la Wikipedia.

Analizando al detalle los resultados de NP-Var, hemos observado que muchos de los fallos de este sistema son producidos debido a errores derivados del preprocesamiento anterior (recuérdese que utilizamos también NP*misc*, ya que pueden corresponderse con NPer mal clasificados). Así, además de algún error de reconocimiento de nombres propios (p. e., "La Fílida de Gálvez de Montalvo" fue dividido en "La" / "Fílida_de_Gálvez_de_Montalvo", siendo esta última entidad marcada como correferencial de "Gálvez de Montalvo"), el NEC etiquetó en repetidas ocasiones como *localidad* u *organización* a NPer ("Chacón": organización, "Santos": localidad, "Gómez": localidad, etc.).

Por lo tanto, y con el fin de conocer el funcionamiento del sistema asumiendo un *input* óptimo, corregimos manualmente los errores del preprocesamiento, obteniendo los resultados de la Tabla 3.

Teniendo en cuenta que el preprocesamiento *nunca* será perfecto, hemos realizado algunas pruebas añadiendo al algoritmo reglas de modificación de la clasificación de las entidades si estas coincidiesen con los apellidos de una entidad ya reconocida. Así, la forma "Lorca" sería siempre etiquetada como persona (aunque anteriormente fuese clasificada como localidad), si "Federico García Lorca" fuese un nombre común en el texto analizado. A pesar de que los resultados preliminares parecen positivos, es necesario realizar más evaluaciones en diferentes *corpora*, para verificar la coexistencia de nombres que, en un mismo documento, se refieran a localidades y a personas.

Los diminutivos son un caso similar: algunas pruebas muestran que el uso de reglas básicas de lematización de diminutivos mejoran su análisis, aunque es necesario evaluar si pueden producir también falsos positivos.

6. Conclusiones y trabajo futuro

En el presente artículo hemos descrito los principales tipos de correferencia entre nombres de persona que existen en documentos periodísticos y enciclopédicos en castellano, así como algunas de las dificultades que surgen a la hora de analizarlos.

Además, se ha definido un algoritmo que, utilizando un conjunto reducido y fácilmente obtenible de recursos externos, resuelve satisfactoriamente la mayoría de estos casos.

Su aplicación previa en sistemas de extracción de información permite unificar los diferentes nombres de persona que se refieren a un mismo individuo. De este modo, el número de oraciones con información biográfica potencialmente relevante aumenta, facilitando al mismo tiempo posteriores procesos de fusión de información extraída de diferentes fuentes.

El sistema no requiere *corpus* de entrenamiento y se compone de reglas independientes del idioma, por lo que la adaptación a otras lenguas próximas es simple. En este sentido, están siendo realizados experimentos en *corpora* portugueses y gallegos.

Como trabajo futuro, además de evaluar el uso del algoritmo como método de corrección de la clasificación semántica, pretendemos ampliar este sistema con un *baseline* de resolución de anáfora de nombres de persona, teniendo como objetivo aumentar el *recall* de sistemas de extracción de información.

Bibliografía

Bontcheva, Kalina, Marin Dimitrov, Diana Maynard, Valentin Tablan, y Hamish Cunningham. 2002. Shallow Methods for Named Entity Coreference Resolution. En *TALN 2002*.

Carreras, Xavier, Lluís Màrquez, y Lluís Padró. 2003. A simple named entity extractor using adaboost. En *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, volumen 4 de *CONLL '03*, páginas 152–155, Stroudsburg, PA, USA. ACL.

Cohen, William W., Pradeep Ravikumar, y Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. En *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*, páginas 73–78.

Fleischman, Michael y Eduard Hovy. 2002. Fine grained classification of named entities. En *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, páginas 1–7, Stroudsburg, PA, USA. ACL.

Kim, Jong-Sun y Martha W. Evens. 1996. Efficient coreference resolution for proper names in the Wall Street Journal Text. En *Online Proceedings of MAICS'96*, Bloomington.

Mani, Inderjeet, Richard T. MacMillan, Susann Luperfoy, Elaine Lusher, y Sharon Laskowski. 1993. Identifying unknown proper names in newswire text. En *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, páginas 44–54, Columbus, Ohio. Special Interest Group on the Lexicon of the ACL.

Mann, Gideon S. 2002. Fine-Grained Proper Noun Ontologies for Question Answering. En *SemaNet'02: Building and Using Semantic Networks*, Taipei, Taiwan.

Mann, Gideon S. y David Yarowsky. 2003. Unsupervised personal name disambiguation. En *Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003*, CONLL '03, páginas 33–40, Stroudsburg, PA, USA. ACL.

Mitkov, Ruslan, Richard Evans, Constantin Orăsan, Cătălina Barbu, Lisa Jones, y Violeta Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. En *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, páginas 49–58, Lancaster, UK.

Padró, Lluís, Miquel Collado, Samuel Reese, Marina Lloberes, y Irene Castellón. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valletta, Malta. ELRA.

Recasens, Marta y Eduard Hovy. 2010. Coreference resolution across corpora: languages, coding schemes, and preprocessing information. En *Proceedings of the 48th Annual Meeting of the ACL*, ACL '10, páginas 1423–1432, Stroudsburg, PA, USA. ACL.

Recasens, Marta y Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*. Available on CJO 2006.

Suchanek, Fabian M., Georgiana Ifrim, y Gerhard Weikum. 2006. LEILA: Learning to Extract Information by Linguistic Analysis. En *Second Workshop on Ontology Population (OLP2) at ACL/COLING*.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, y Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. En *Proceedings of the 6th conference on Message understanding*, MUC6 '95, páginas 45–52, Stroudsburg, PA, USA. ACL.

Wacholder, Nina, Yael Ravin, y Misook Choi. 1997. Disambiguation of proper names in text. En *Proceedings of the fifth conference on Applied natural language processing*, ANLC '97, páginas 202–208, Stroudsburg, PA, USA. ACL.

Wakao, Takahiro, Robert Gaizauskas, y Yorick Wilks. 1996. Evaluation of an algorithm for the recognition and classification of proper names. En *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, páginas 418–423, Stroudsburg, PA, USA. ACL.

Wan, Xiaojun, Jianfeng Gao, Mu Li, y Binggong Ding. 2005. Person resolution in person search results: Webhawk. En *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, páginas 163–170, New York, NY, USA. ACM.